10.12871/CLICIT2014162

Developing corpora and tools for sentiment analysis: the experience of the University of Turin group

Manuela Sanguinetti, Emilio Sulis, Viviana Patti, Giancarlo Ruffo, Leonardo Allisio, Valeria Mussa and Cristina Bosco
Dipartimento di Informatica
Università di Torino
{msanguin,sulis,patti,ruffo,bosco@di.unito.it},
{leonardo.allisio,valeria.mussa@studenti.unito.it}

Abstract

English. The paper describes the ongoing experience at the University of Turin in developing linguistic resources and tools for sentiment analysis of social media. We describe in particular the development of Senti-TUT, a human annotated corpus of Italian Tweets including labels for sentiment polarity and irony, which has been recently exploited within the SENTIment POLarity Classification shared task at Evalita 2014. Furthermore, we report about our ongoing work on the Felicittà web-based platform for estimating happiness in Italian cities, which provides visualization techniques to interactively explore the results of sentiment analysis performed over Italian geotagged Tweets.

Italiano. *L'articolo presenta l'esperienza fatta presso l'Università di Torino nello sviluppo di risorse linguistiche e strumenti per la sentiment analysis di social media. In particolare, viene descritto Senti-TUT, un corpus di Tweet in Italiano, che include annotazioni relative alla polarità del sentiment e alla presenza di ironia, utilizzato nell'ambito del task di SENTIment POLarity Classification di Evalita 2014. Inoltre viene presentato il lavoro su Felicittà, una piattaforma Web per la stima della felicità nelle città italiane, che fornisce diverse modalità di visualizzazione del grado di felicità che emerge da un'analisi del sentiment su messaggi Twitter geolocalizzati in Italiano.*

1 Introduction

Several efforts are currently devoted to automatically mining opinions and sentiments from natural language, e.g. in social media posts, news and reviews about commercial products. This task entails a deep understanding of the explicit and implicit information conveyed by the language, and most of the approaches applied refer to annotated corpora and adequate tools for their analysis.

In this paper, we will describe the experiences carried on at the Computer Science Department of the University of Turin in the development of corpora and tools for Sentiment Analysis and Opinion Mining (SA&OM) during the last few years. These experiences grew and are still growing in a scenario where an heterogeneous group of researchers featured by skills varying from computational linguistics, sociology, visualization techniques, big data analysis and ontologies cooperates. Both the annotation applied in the developed corpora and the tools for analyzing and displaying data analysis depend in fact on a cross-fertilization of different research areas and on the expertise gained by the group members in their respective research fields. The projects we will describe are currently oriented to the investigation of aspects of data analysis that can be observed in such a particular perspective, e.g. figurative language or disagreement deep analysis, rather than to the achievement of high scores in the application of classifiers and statistical tools.

The paper is organized as follows. The next section provides an overview on the annotated corpus Senti-TUT, which includes two main datasets: TW-NEWS (political domain) and TW-FELICITTA (generic collection), while Section 3 describes the main uses of Senti-TUT and the Felicittà application context.

2 Annotating corpora for SA&OM

The experience on human annotation of social media data for SA&OM mainly refers to the Senti-TUT corpus of Italian Tweets, featured by different stages of development (Gianti et al., 2012; Bosco et al., 2013; Bosco et al., 2014). We have

relied on our skills in building linguistic resources, such as TUT[1].

Tweets have been annotated at the message level. Among the main goals we pursued in the annotation of this corpus, there is the study of irony, a specific phenomenon which can affect SA&OM systems performances (Riloff et al., 2013; Reyes et al., 2012; Reyes et al., 2013; Hao and Veale, 2010; González-Ibáñez et al., 2011; Davidov et al., 2011; Maynard and Greenwood, 2014; Rosenthal et al., 2014). To deal with this issue, we extended a traditional polarity-based framework with a new dimension which explicitly accounts for irony. According to literature, boundaries in meaning between different types of irony are fuzzy (Gibbs and Colston, 2007) and this could be an argument in favor of annotation approaches where different types of irony are not distinguished, as the one adopted in Senti-TUT. We thus designed and applied to the collected data an annotation oriented to the description of Tweet polarity, which is suitable for high level tasks, such as classifying the polarity of a given text. The annotation scheme included the traditional labels for distinguishing among positive, negative or neutral sentiment. Moreover, we introduced the labels HUM, to mark the intention of the author of the post to express irony or sarcasm, and MIXED, to mark the presence of more than one sentiment within a Tweet[2]. Summarizing, our tagset includes:

POS	positive
NEG	negative
NONE	neutral (no sentiment expressed)
MIXED	mixed (POS and NEG both)
HUM	ironic
UN	unintelligible

Having a distinguished tag for irony did not prevent us from reconsidering these Tweets at a later stage, and force their classification according to traditional annotation schemes for the sentiment analysis task, i.e. applying a positive or negative polarity label, e.g. to measure how an automatic traditional sentiment classifier can be wrong, as we did in (Bosco et al., 2013). Similarly, identifying Tweets containing mixed sentiment can be useful in order to measure how the phenomenon impacts the performances of sentiment classifiers. Moreover, having distinguished tags for irony and mixed sentiment can be helpful to a better development of the corpora, in order to increase the inter-annotator agreement, since cases, that typically can be source of disagreement on the polarity valence, are recognized and labeled separately.

2.1 The Senti-TUT core

The first stage of development of the Senti-TUT project [3] led to the results described in (Bosco et al., 2013; Gianti et al., 2012). The major aims of the project are the development of a resource missing for Italian, and the study of a particular linguistic device: irony. This motivated the selection of data domain and source, i.e. politics and Twitter: Tweets expressing political opinions contain extensive use of irony. The corpus developed at this stage includes a dataset called TW-NEWS, composed of $3,288$ posts collected in the time frame between October 2012 and February 2013 and that focuses on the past Monti's government in Italy. They where collected and filtered, relying on the Blogmeter social media monitoring platform[4]. For each post in TW-NEWS, we collected in the first phase two independent annotations. The inter-annotator agreement calculated at this stage, according to the Cohen's κ score, was $\kappa = 0.65$ (Artstein and Poesio, 2008). The second step entailed the collection of cases when the annotators disagreed (about 25% of data). A third annotator thus attempted to solve the disagreement or discarded the inherently disagreement cases (around 2% of the data). This is motivated by the need of datasets that can be sufficiently unambiguous to be useful for training of classifiers and automatic tools. A second dataset, called TW-SPINO and composed of $1,159$ messages from the Twitter section of Spinoza [5] (a very popular Italian blog of posts with sharp satire on politics) has been collected in order to extend the size of the set of ironic Tweets tagged as HUM.

2.2 The TW-FELICITTA corpus

The TW-FELICITTA corpus (Bosco et al., 2014) can be seen as a further extension of Senti–TUT, mainly developed to validate the approach applied

[1] http://www.di.unito.it/~tutreeb

[2] About the MIXED label see also the gold standard presented in (Saif et al., 2013)

[3] http://www.di.unito.it/~tutreeb/sentiTUT.html

[4] http://www.blogmeter.eu

[5] http://www.spinoza.it

in the Felicittà project (see Section 3). The $1,500$ Italian Tweets here collected were randomly extracted from those collected by Twitter API, paying attention at avoiding geographic and temporal bias.

TW-FELICITTA corpus is a general-purpose resource. This means that data are not filtered in some way, but are more representative of the Twitter language and topics in general. The absence of a specific domain context made the interpretation and annotation of the posts more difficult. The annotation process involved four human annotators. We collected not less than three independent annotations for each Tweet according to the annotation scheme described above and relying on a set of shared guidelines. The inter–annotator agreement achieved was 0.51 (Fleiss, 1971). Hypothesizing that the 'soft disagreement' (i.e. disagreement occurring when we detect two agreeing and one disagreeing tags) was at least in part motivated by annotators biases or errors, after a further discussion of the guidelines, we applied a fourth independent annotation to the Tweets in soft disagreement. The resulting final corpus consists of $1,235$ Tweets with agreed annotation and 265 Tweets with disagreed annotation.

Table1 presents an overview of the distribution of tags (UN excluded) referring to the three annotated datasets currently included in Senti-TUT.

label	News	Felicittà	Spino
POS	513	338	-
NEG	788	368	-
NONE	1.026	260	-
MIXED	235	39	-
HUM	726	168	1.159

Table 1: Distribution of Senti-TUT tags in TW-NEWS, TW-FELICITTA and TW-SPINO.

The development of TW-FELICITTA also provided the basis for reflecting on the need of a framework to capture and analyze the nature of the disagreement (i.e. Tweets on which the disagreement reflects semantic ambiguity in the target instances and provides useful information). Hypothesizing that the analysis of the disagreement should be considered as a starting point for a deeper understanding of the task to be automated in our sentiment engine (in tune with the arguments in (Inel et al., 2014)), we investigated the use of different measures to analyze the following complementary aspects: the *subjectivity of each sentiment label* and the *subjectiveness of the involved annotators*.

Agreement analysis For what concerns the detection of the *subjectivity of the sentiment labels* in our annotation scheme, we hypothesized that when a sentiment label is more involved in the occurrence of disagreement, this is because it is more difficult to annotate, as its meaning is less shared among the annotators and there is a larger range of subjectivity in its interpretation. In order to estimate the subjectivity degree of each label *L*, we calculated the percentage of cases where *L* produced an agreement or disagreement among annotators. Table 2 shows how much a label has been used in percentage to contribute to the definition of an agreed or disagreed annotation of the Tweets.

label	agreement	disagreement
POS	26.3	14.4
NEG	29.2	17.8
NONE	21.8	23.5
MIXED	3.3	8.8
HUM	11.9	13.0
UN	7.6	22.5

Table 2: A measure of subjectivity of tags annotated in TW-FELICITTA

It should be observed, in particular, that while POS and NEG labels seem to have a higher reference to the agreement, for UN and MIXED the opposite situation happens.

Assuming a perspective oriented to the single annotators and referring to all the annotated tags, as above, we also measured the *subjectiveness* of each *annotator involved in the task* according to the variation in the exploitation of the labels. For each label *L*, starting from the total amount of times when *L* has been annotated, we calculated the average usage of the label. Then we calculated the deviation with respect to the average and we observed how this varies among the annotators. The deviation with respect to the average usage of the label is maximum for the MIXED and UN tags, and minimum for POS and NEG, showing that the annotators are more confident in exploit-

ing the latter tags (Table 3).

label	total	avg	dev. +	dev. -
NEG	1,592	398	15.32%	14.82%
POS	1,421	355.25	6.68%	5.13%
NONE	1,281	320.25	24.90%	16.31%
HUM	700	175	28.57%	31.42%
UN	569	142	73.94%	35.21%
MIXED	237	59.25	46.83%	80.18%

Table 3: A measure of variation among the exploitation of the labels in TW-FELICITTA.

3 Exploitation of Senti-TUT and ongoing applications of SA on Italian tweets

Irony and emotion detection A preliminary corpus-based analysis of phenomena connected to irony, in particular polarity reversing and frequency of emotions, is reported in (Bosco et al., 2013) and involved the Tweets tagged by HUM in TW-NEWS and TW-SPINO. We applied rule-based automatic classification techniques in (Bolioli et al., 2013) to annotate ironic Tweets according to seven categories: Ekman's basic emotions (*anger, disgust, fear, joy, sadness, surprise*) plus *love*. These emotions were expressed in 20% of the dataset and distributed differently in the corpora. What emerged was that irony was often used in conjunction with a seemingly positive statement to reflect a negative one (rarely the other way around). This is in accordance with theoretical accounts (Gibbs and Colston, 2007), reporting that expressing a positive attitude in a negative mode is rare and harder for humans to process, as compared to expressing a negative attitude in a positive mode.

Felicittà Felicittà (Allisio et al., 2013) is an online platform for estimating happiness in the Italian cities, which daily analyzes Twitter posts and exploits temporal and geo-spatial information related to Tweets, in order to enable the summarization of SA outcomes. The system automatically analyzes posts and classifies them according to traditional polarity labels according to a pipeline which performs a shallow analysis of Tweets and applies a lexicon-based approach looking for the word polarity in WordNetAffect (Strapparava and Valitutti, 2004). At the current stage of the project, we are investigating both the visualization techniques and data aggregation, also in order to compare, in future works, results extracted from Twitter about specific topics to those extracted from other sources like demographic reports.

SENTIPOLC For what concerns the exploitation of the Senti–TUT corpus, a further step is related to its use within the new shared task on sentiment analysis in Twitter (SENTIment POLarity Classification – SENTIPOLC[6]), as part of Evalita 2014[7]. SENTIPOLC represents a valuable forum to validate the data and to compare our experience to that of both the participants and colleagues co-organizing the task from University of Bologna, University of Groningen, Universitat Politècnica de València and the industry partner Blogmeter (CELI). (Basile et al., 2014). The main focus is on detecting sentiment polarity in Twitter messages as in SemEval 2013 - Task 2 (Nakov et al., 2013), but a pilot task on irony detection has been also proposed. The datasets released include Twitter posts from TW–NEWS and TW–FELICITTA annotated by the Turin & Blogmeter teams, and other posts collected and annotated by Bologna, Groningen and València teams (Basile and Nissim, 2013).

4 Conclusion

The paper describes the experiences done at the University of Turin on topics related to SA&OM, with a special focus on the main directions we are following. The first one is the development of annotated corpora for Italian that can be exploited both in automatic systems' training, in evaluation fora, and in investigating the nature of the data itself, also by a detailed analysis of the disagreement occurring in the datasets. The second direction, which is exemplified by ongoing work on the Felicittà platform, consists in the development of applications of SA on social media in the social and behavioral sciences field, where SA techniques can contribute to interpret the degree of well-being of a country (Mitchell et al., 2013; Quercia et al., 2012), with a special focus on displaying the results generated by the analysis in a graphic form that can be easily readable also for non-expert users.

[6] http://www.di.unito.it/~tutreeb/sentipolc-evalita14/

[7] http://www.evalita.it

Acknowledgments

We acknowledge Ing. Sergio Rabellino, who leads the ICT team of the Computer Science Department at the University of Turin, for the support in the development of a web platform for Twitter annotated data release based on RESTful technology.

References

Leonardo Allisio, Valeria Mussa, Cristina Bosco, Viviana Patti, and Giancarlo Ruffo. 2013. Felicittà: Visualizing and estimating happiness in Italian cities from geotagged Tweets. In *Proc. of the 1st Int. Workshop on Emotion and Sentiment in Social and Expressive Media, ESSEM@AI*IA*, volume 1096, pages 95–106. CEUR-WS.org.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, December.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107, Atlanta.

Valerio Basile, Andrea Bolioli, Malvina Nissim, Viviana Patti, and Paolo Rosso. 2014. Overview of the Evalita 2014 SENTIment POLarity Classification Task. In *Proceedings of the 4th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'14)*, Pisa, Italy.

Andrea Bolioli, Federica Salamino, and Veronica Porzionato. 2013. Social media monitoring in real life with blogmeter platform. In *ESSEM@AI*IA*, volume 1096 of *CEUR Workshop Proceedings*, pages 156–163. CEUR-WS.org.

Cristina Bosco, Viviana Patti, and Andrea Bolioli. 2013. Developing corpora for sentiment analysis: The case of irony and Senti-TUT. *IEEE Intelligent Systems*, 28(2):55–63.

Cristina Bosco, Leonardo Allisio, Valeria Mussa, Viviana Patti, Giancarlo Ruffo, Manuela Sanguinetti, and Emilio Sulis. 2014. Detecting happiness in italian tweets: Towards an evaluation dataset for sentiment analysis in felicittà. In B. Schuller, P. Buitelaar, L. Devillers, C. Pelachaud, T. Declerck, A. Batliner, P. Rosso, and S. Gaines, editors, *Proc. of the 5th International Workshop on Emotion, Social Signals, Sentiment and Linked Opena Data, ESSSLOD 2014*, pages 56–63, Reykjavik, Iceland. European Language Resources Association.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2011. Semi-supervised recognition of sarcastic sentences in Twitter and Amazon. In *Proceedings of the CONLL'11*, pages 107–116, Portland, Oregon (USA).

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Andrea Gianti, Cristina Bosco, Viviana Patti, Andrea Bolioli, and Luigi Di Caro. 2012. Annotating irony in a novel italian corpus for sentiment analysis. In *Proceedings of the 4th Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3@LREC'12)*, pages 1–7, Istanbul, Turkey.

Raymond W Gibbs and Herbert L. Colston, editors. 2007. *Irony in Language and Thought*. Lawrence Erlbaum Associates, New York.

Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. 2011. Identifying sarcasm in twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 581–586, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yanfen Hao and Tony Veale. 2010. An ironic fist in a velvet glove: Creative mis-representation in the construction of ironic similes. *Minds Mach.*, 20(4):635–650, November.

Oana Inel, Khalid Khamkham, Tatiana Cristea, Anca Dumitrache, Arne Rutjes, Jelle van der Ploeg, Lukasz Romaszko, Lora Aroyo, and Robert-Jan Sips. 2014. Crowdtruth: Machine-human computation framework for harnessing disagreement in gathering annotated data. In Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig Knoblock, Denny Vrandecic, Paul Groth, Natasha Noy, Krzysztof Janowicz, and Carole Goble, editors, *Proceedings of ISWC 2014*, volume 8797 of *Lecture Notes in Computer Science*, pages 486–504. Springer International Publishing.

Diana Maynard and Mark Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association.

Lewis Mitchell, Morgan R. Frank, Kameron Decker Harris, Peter Sheridan Dodds, and Christopher M. Danforth. 2013. The geography of happiness: Connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLoS ONE*, 8(5), 05.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on*

Semantic Evaluation (SemEval 2013), pages 312–320, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Daniele Quercia, Jonathan Ellis, Licia Capra, and Jon Crowcroft. 2012. Tracking "gross community happiness" from tweets. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, pages 965–968, New York, NY, USA. ACM.

Antonio Reyes, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data Knowledge Engineering*, 74:1–12.

Antonio Reyes, Paolo Rosso, and Tony Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, 47(1):239–268.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proc. of EMNLP*, pages 704–714. ACL.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Hassan Saif, Miriam Fernandez, He Yulan, and Harith Alani. 2013. Evaluation datasets for twitter sentiment analysis: A survey and a new dataset, the STS-Gold. In *Proc. of the 1st Int. Workshop on Emotion and Sentiment in Social and Expressive Media, ESSEM@AI*IA*, volume 1096 of *CEUR Workshop Proceedings*, pages 9–21. CEUR-WS.org.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: an affective extension of WordNet. In *Proc. of the 4th Language Resources and evaluation Conference, LREC'04*, volume 4, pages 1083–1086. ELRA.